REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
23-11-2015	Final Report	1-Oct-2011 - 30-Jun-2015

4. TITLE AND SUBTITLE
Final Report: The Human Microbiome as a Multipurpose Biomarker

5a. CONTRACT NUMBER
W911NF-11-1-0473

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Curtis Huttenhower

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Harvard School of Public Health
Biostatistics
President and Fellows of Harvard College
Boston, MA 02115 -6028

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
60287-MA.22

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
The human microbiome comprises the communities of microbes carried in and on the body in health and disease, including trillions of bacteria, viruses, archaea, and fungi per individual. These microbiota, which defend us against pathogens and help digest our food, are personalized among individuals. The specific microbes present at any one habitat within an individual become relatively stable during the first several years of life, but change in as-yet-uncharacterized ways as a host is exposed to new environments, diets, locations, and social contacts. The microbial composition of a given individual might thus be linked to his genetic background or early life history, for example

15. SUBJECT TERMS
microbiome, biomarker, microbial forensics, microbial ecology, identifiability

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Curtis Huttenhower
UU	UU	UU	UU		19b. TELEPHONE NUMBER 617-432-4912

Standard Form 298 (Rev 8/98)
Prescribed by ANSI Std. Z39.18

Report Title

Final Report: The Human Microbiome as a Multipurpose Biomarker

ABSTRACT

The human microbiome comprises the communities of microbes carried in and on the body in health and disease, including trillions of bacteria, viruses, archaea, and fungi per individual. These microbiota, which defend us against pathogens and help digest our food, are personalized among individuals. The specific microbes present at any one habitat within an individual become relatively stable during the first several years of life, but change in as-yet-uncharacterized ways as a host is exposed to new environments, diets, locations, and social contacts. The microbial composition of a given individual might thus be linked to his genetic background or early life history, for example, while the metabolism of those microbes would reveal more about his recent medical history or diet. The goals of this project are thus 1) to assess the structure of any microbial habitat and its potential for identifiability (including dietary history, medical history, biometric, demographics or environmental exposures) and 2) to determine the relationships and interactions between the microbial communities within a host.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received		Paper
07/30/2012	2.00	Karoline Faust, J. Fah Sathirapongsasuti, Jacques Izard, Nicola Segata, Dirk Gevers, Jeroen Raes, Curtis Huttenhower, Christos A. Ouzounis. Microbial Co-occurrence Relationships in the Human Microbiome, PLoS Computational Biology, (07 2012): 0. doi: 10.1371/journal.pcbi.1002606
07/30/2012	1.00	Dirk Gevers, Curtis Huttenhower, Nicola Segata, Susan Haake, Peter Mannon, Katherine P Lemon, Levi Waldron, Jacques Izard. Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples, Genome Biology and Evolution, (06 2012): 0. doi: 10.1186/gb-2012-13-6-r42
08/30/2013	4.00	Xochitl C Morgan, Timothy L Tickle, Harry Sokol, Dirk Gevers, Kathryn L Devaney, Doyle V Ward, Joshua A Reyes, Samir A Shah, Neal LeLeiko, Scott B Snapper, Athos Bousvaros, Joshua Korzenik, Bruce E Sands, Ramnik J Xavier, Curtis Huttenhower. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment, Genome Biology, (04 2012): 0. doi: 10.1186/gb-2012-13-9-r79
10/09/2014	5.00	E. A. Franzosa, X. C. Morgan, N. Segata, L. Waldron, J. Reyes, A. M. Earl, G. Giannoukos, M. R. Boylan, D. Ciulla, D. Gevers, J. Izard, W. S. Garrett, A. T. Chan, C. Huttenhower. Relating the metatranscriptome and metagenome of the human gut, Proceedings of the National Academy of Sciences, (05 2014): 0. doi: 10.1073/pnas.1319284111
10/09/2014	6.00	Afrah Shafquat, Regina Joice, Sheri L. Simmons, Curtis Huttenhower. Functional and phylogenetic assembly of microbial communities in the human microbiome, trends in microbiology, (05 2014): 0. doi: 10.1016/j.tim.2014.01.011
10/09/2014	7.00	Subra Kugathasan, Lee A. Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, Dan Knights, Se Jin Song, Moran Yassour, Xochitl C. Morgan, Aleksandar D. Kostic, Chengwei Luo, Antonio González, Daniel McDonald, Dirk Gevers, Yael Haberman, Thomas Walters, Susan Baker, Joel Rosh, Michael Stephens, Melvin Heyman, James Markowitz, Robert Baldassano, Anne Griffiths, Francisco Sylvester, David Mack, Sandra Kim, Wallace Crandall, Jeffrey Hyams, Curtis Huttenhower, Rob Knight, Ramnik J. Xavier. The Treatment-Naive Microbiome in New-Onset Crohn's Disease, Cell Host & Microbe, (03 2014): 0. doi: 10.1016/j.chom.2014.02.005
TOTAL:		**6**

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received		Paper
11/19/2015	9.00	Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B6, Schwager E, Knights D, Song SJ, Yassour M, Morgan XC, Kostic AD, Luo C, González A, McDonald D, Haberman Y, Walters T, Baker S, Rosh J, Stephens M, Heyman M, Markowitz J, Baldassano R, Griffiths A, Sylvester F, Mack D, Kim S, Crandall W, Hyams J, Huttenhower C, Knight R, Xavier RJ. The treatment-naive microbiome in new-onset Crohn's disease, Cell Host & Microbe, (03 2014): 382. doi:
11/19/2015	18.00	Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C , Segata N. MetaPhlAn2 for enhanced metagenomic taxonomic profiling, Nature Methods, (09 2015): 902. doi:
11/19/2015	17.00	Silverberg MS, Weersma RK, Gevers D, Dijkstra G, Huang H, Tyler AD, van Sommeren S, Imhann F, Knights D, Stempak JM, Huang H, Vangay P, Al-Ghalith GA, Russell C, Sauk J, Knight J, Daly MJ, Huttenhower C, Xavier RJ. Complex host genetics influence the microbiome in inflammatory bowel disease, Genome Medcine, (12 2014): 107. doi:
11/19/2015	10.00	Morgan XC1, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, Reyes JA, Shah SA, LeLeiko N, Snapper SB, Bousvaros A, Korzenik J, Sands BE, Xavier RJ, Huttenhower C. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment, Genome Biology, (04 2012): 79. doi:
11/19/2015	11.00	Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn, PeerJ, (06 2015): 1029. doi:
11/19/2015	12.00	Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, Huttenhower C. Microbial co-occurrence relationships in the human microbiome, PLoS Computational Biology, (07 2012): 1002606. doi:
11/19/2015	13.00	Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower C. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences, Nature Biotechnology, (09 2013): 814. doi:
11/19/2015	14.00	Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, Giannoukos G, Boylan MR, Ciulla D, Gevers D, Izard J, Garrett WS, Chan AT, Huttenhower C. Relating the metatranscriptome and metagenome of the human gut, proceeding of the national academy of sciences, (06 2014): 2329. doi:
11/19/2015	15.00	Segata N, Haake SK, Mannon P, Lemon KP, Waldron L, Gevers D, Huttenhower C, Izard J. Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples, Genome Biology, (06 2012): 42. doi:
11/20/2015	16.00	Gevers D, Knight R, Petrosino JF, Huang K, McGuire AL, Birren BW, Nelson KE, White O, Methé BA, Huttenhower C. The Human Microbiome Project: a community resource for the healthy human microbiome, A Peer-Reiewed Open Access, (08 2012): 1001377. doi:

11/20/2015 19.00 Lewis CM, Lewis L, Ley RE, Li K, Liolios K, Liu B, Liu Y, Lo CC, Lozupone CA, Lunsford R, Madden T, Mahurkar AA, Mannon PJ, Mardis ER, Markowitz VM, Mavrommatis K, McCorrison JM, McDonald D, McEwen J, McGuire AL, McInnes P, Mehta T, Mihindukulasuriya KA, Miller JR, Minx PJ, Newsham I, Nusbaum C, O'Laughlin M, Orvis J, Pagani I, Palaniappan K, Patel SM, Pearson M, Peterson J, Podar M, Pohl C, Pollard KS, Priest ME, Proctor LM, Qin X, Raes J, Ravel J, Reid JG, Rho M, Rhodes R, Riehle KP, Rivera MC, Rodriguez-Mueller B, Rogers YH, Ross MC, Russ C, Sanka RK, Sankar P, Sathirapongsasuti J, Schloss JA, Schloss PD, Schmidt TM, Scholz M, Schriml L, Schubert AM, Segata N, Segre JA, Shannon WD, Sharp RR, Sharpton TJ, Shenoy N, Sheth NU, Simone GA, Singh I, Smillie CS, Sobel JD, Sommer DD, Spicer P, Sutton GG, Sykes SM, Tabbaa DG, Thiagarajan M, Tomlinson CM, Torralba M, Treangen TJ, Truty RM, Vishnivetskaya TA, Walker J, Wang L, Wang Z, Ward DV, Warren W, Watson MA, Wellington C, Wetterstrand KA, White JR, Wilczek-Boney K, Wu YQ, Wylie KM, Wylie T, Yandava C, Ye L, Ye Y, Yooseph S, Youmans BP, Zhang L, Zhou Y, Zhu Y, Zoloth L, Zucker JD, Birren BW, Gibbs RA, Highlander SK, Weinstock GM, Wilson RK, White O, Methé BA, Nelson KE, Pop M, Creasy HH, Giglio MG, Huttenhower C, Gevers D, Petrosino JF, Abubucker S, Badger JH, Chinwalla AT, Earl AM, FitzGerald MG, Fulton RS, Hallsworth-Pepin K, Lobos EA, Madupu R, Magrini V, Martin JC, Mitreva M, Muzny DM, Sodergren EJ, Versalovic J, Wollam AM, Worley KC, Wortman JR, Young SK, Zeng Q, Aagaard KM, Abolude OO, Allen-Vercoe E, Alm EJ, Alvarado L, Andersen GL, Anderson S, Appelbaum E, Arachchi HM, Armitage G, Arze CA, Ayvaz T, Baker CC, Begg L, Belachew T, Bhonagiri V, Bihan M, Blaser MJ, Bloom T, Bonazzi VR, Brooks P, Buck GA, Buhay CJ, Busam DA, Campbell JL, Canon SR, Cantarel BL, Chain PS, Chen IM, Chen L, Chhibba S, Chu K, Ciulla DM, Clemente JC, Clifton SW, Conlan S, Crabtree J, Cutting MA, Davidovics NJ, Davis CC, DeSantis TZ, Deal C, Delehaunty KD, Dewhirst FE, Deych E, Ding Y, Dooling DJ, Dugan SP, Dunne W Jr, Durkin A, Edgar RC, Erlich RL, Farmer CN, Farrell RM, Faust , Feldgarden M, Felix VM, Fisher S, Fodor AA, Forney L, Foster L, Di Francesco V, Friedman J, Friedrich DC, Fronick CC, Fulton LL, Gao H, Garcia N, Giannoukos G, Giblin C, Giovanni MY, Goldberg JM, Goll J, Gonzalez A, Griggs A, Gujja S, Haas BJ, Hamilton HA, Harris EL, Hepburn TA, Herter B, Hoffmann DE, Holder ME, Howarth C, Huang KH, Huse SM, Izard J, Jansson JK, Jiang H, Jordan C, Joshi V, Katancik JA, Keitel WA, Kelley ST, Kells C, Kinder-Haake S, King NB, Knight R, Knights D, Kong HH, Koren O, Koren S, Kota KC, Kovar CL, Kyrpides NC, La Rosa PS, Lee SL, Lemon KP, Lennon N. A framework for human microbiome research,
Nature, (06 2012): 215. doi:

11/20/2015 20.00 Versalovic J, Wollam AM, Worley KC, Wortman JR, Young SK, Zeng Q, Aagaard KM, Abolude OO, Allen-Vercoe E, Alm EJ, Alvarado L, Andersen GL, Anderson S, Schloss PD, Schmidt TM, Scholz M, Schriml L, Schubert AM, Segata N, Segre JA, Shannon WD, Sharp RR, Sharpton TJ, Shenoy N, Sheth NU, Simone GA, Singh I, Smillie CS, Sobel JD, Sommer DD, Spicer P, Sutton GG, Sykes SM, Tabbaa DG, Thiagarajan M, Tomlinson CM, Torralba M, Treangen TJ, Truty RM, Vishnivetskaya TA, Walker J, Wang L, Wang Z, Ward DV, Warren W, Watson MA, Wellington C, Wetterstrand KA, White JR, Wilczek-Boney K, Wu Y, Wylie KM, Wylie T, Yandava C, Ye L, Ye Y, Yooseph S, Youmans BP, Zhang L, Zhou Y, Zhu Y, Zoloth L, Zucker JD, Birren BW, Gibbs RA, Highlander SK, Methé BA, Nelson KE, Petrosino JF, Weinstock GM, Wilson RK, White O, Madupu R, Magrini V, Martin JC, Mitreva M, Muzny DM, Sodergren EJ, Qin X, Raes J, Ravel J, Reid JG, Rho M, Rhodes R, Riehle KP, Rivera MC, Rodriguez-Mueller B, Rogers YH, Ross MC, Russ C, Sanka RK, Sankar P, Sathirapongsasuti J, Schloss JA, Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS, Giglio MG, Hallsworth-Pepin K, Lobos EA, Blaser MJ, Bloom T, Bonazzi V, Brooks J, Appelbaum E, Buck GA, Arachchi HM, Armitage G, Arze CA, Ayvaz T, Baker CC, Begg L, Belachew T, Bhonagiri V, Bihan M, Buhay CJ, Busam DA, Campbell JL, Canon SR, Cantarel BL, Chain PS, Chen IM, Chen L, Chhibba S, Chu K, Ciulla DM, Clemente JC, Clifton SW, Conlan S, Crabtree J, Cutting MA, Davidovics NJ, Davis CC, DeSantis TZ, Deal C, Delehaunty KD, Dewhirst FE, Deych E, Ding Y, Dooling DJ, Dugan SP, Dunne WM, Durkin A, Edgar RC, Erlich RL, Farmer CN, Farrell RM, Faust K, Feldgarden M, Felix VM, Fisher S, Fodor AA, Forney LJ, Foster L, Di Francesco V, Friedman J, Friedrich DC, Fronick CC, Fulton LL, Gao H, Garcia N, Giannoukos G, Giblin C, Giovanni MY, Goldberg JM, Goll J, Gonzalez A, Griggs A, Gujja S, Haake SK, Haas BJ, Hamilton HA, Harris E, Hepburn TA, Herter B, Hoffmann DE, Holder ME, Howarth C, Huang KH, Huse SM, Izard J, Jansson JK, Jiang H, Jordan C, Joshi V, Katancik JA, Keitel WA, Kelley SL, Kells C, King NB, Knights D, Kong HH, Koren O, Koren S, Kota KC, Kovar CL, Kyrpides NC, La Rosa PS, Lee SL, Lemon KP, Lennon N, Lewis CM, Lewis L, Ley RE, Li K, Liolios K, Liu B, Liu Y, Lo CC, Lozupone CA, Lunsford R, Madden T, Mahurkar AA, Mannon PJ, Mardis ER, Markowitz VM, Mavromatis K, McCorrison JM, McDonald D, McEwen J, McGuire AL, McInnes P, Mehta T, Mihindukulasuriya KA, Miller JR, Minx PJ, ewsham I, Nusbaum C, O'Laughlin M, Orvis J, Pagani I, Palaniappan K, Patel SM, Pearson M, Peterson J, Podar M, Pohl C, Pollard KS, Pop M, Priest ME, Proctor LM. Structure, function and diversity of the healthy human microbiome,
Nature, (06 2012): 207. doi:

11/20/2015	21.00	Huttenhower C, Knight R, Brown CT, Caporaso JG, Clemente JC, Gevers D, Franzosa EA, Kelley ST, Knights D, Ley RE, Mahurkar A, Ravel J, Scientists for Advancement of Microbiome Research, White O. Advancing the microbiome research community, Cell, (10 2014): 227. doi:

TOTAL: **13**

Number of Papers published in non peer-reviewed journals:

(c) Presentations

1- "High-precision functional profiling of microbial communities and the human microbiome." Wellcome Trust Workshop on Applied Bioinformatics and Public Health Microbiology. Hinxton, UK, 2015
2- "High-precision functional profiling of microbial communities and the human microbiome." Wellcome Trust Workshop on Applied Bioinformatics and Public Health Microbiology. Hinxton, UK, 2015
3- "High-precision functional profiling of microbial communities and the human microbiome." Canadian Institute for Health Research STAGE seminar. Toronto, Canada, 2015
"Towards systems-level functional profiling of microbial communities and the human microbiome." University of Pennsylvania Microbiology seminar. Philadelphia, PA, 2015
4- "High-precision Functional Profiling of Microbial Communities and the Human Microbiome." 41st Annual Northeast Bioengineering Conference. Albany, NY, 2015
5- "High-precision functional profiling and integration of metagenomes and metatranscriptomes." International Human Microbiome Congress workshop on Integrated 'Omics for Microbiome Analyses. Luxembourg, Luxembourg, 2015
6- "High-precision functional profiling of microbial communities and the human microbiome." Simons Foundation Symposium on Genomics in Single Cells and Microbiomes. New York, NY, 2015
7- "A Tour of the bioBakery: Computational Tools for Microbial Community Analysis." Broad Institute Medical and Population Genetics seminar. Cambridge, MA, 2015 (presented by Eric Franzosa)
8- "Towards systems-level functional profiling of microbial communities and the human microbiome." Channing Division of Network Medicine Theodore L. Badger Lecture. Boston, MA, 2015
9- "The microbiome in IBD and analysis methods for microbial communities." International Inflammatory Bowel Disease Genetics Consortium meeting. Barcelona, Spain, 2015
10- "An Introduction to Microbial Community Analyses." Evomics and Genomics workshop. Cesky Krumlov, Czech Republic, 2015
11- "High-specificity methods for profiling microbial communities and the human microbiome." University of Oregon Computer Science colloquium. Eugene, OR, 2014
12- "Metagenomics, metatranscriptomics, and multi'omic integration." Massachusetts General Hospital Center for the Study of Inflammatory Bowel Disease research symposium. Boston, MA, 2014
13- "High-precision methods for metagenomic and metatranscriptomic profiling." New York University Medical School seminar. New York, NY, 2014
14- "High-precision profiling of microbial communities and the human microbiome." University of Oregon Institute for Theoretical Sciences seminar. Eugene, OR, 2014
15- "Computational Approaches for the Human Microbiome in Health and Disease," 12th Biennial Congress of the Anaerobe Society of the Americas. Chicago, IL, 2014
16- "An introduction to the microbiome and quantitative methods for microbial community analysis," HSPH Biostatistics Summer Program in Quantitative Sciences. Boston, MA, 2014
17- "An introduction to the microbiome and methods for microbial community analysis," Harvard/MIT Minority Introduction to Engineering, and Science. Boston, MA, 2014
18- "An introduction to metagenomics," Strategies and Techniques for Analyzing Microbial Population Structure. Woods Hole, MA, 2014
19- "Identifiability of the Human Microbiome," Dalhousie University Centre for Comparative Genomics and Evolutionary Bioinformatics and Microbiome User Group. Halifax, Canada, 2014
20- "A Tour of the BioBakery: Computational Tools for Microbial Community Analysis," Harvard School of Public Health Program in Quantitative Genomics Short Course series. Boston, MA, 2014 (presented by Eric Franzosa)
21- "High-precision functional profiling and integration of metagenomes and metatranscriptomes," Weizmann Institute Systems Biology Seminar Series. Rehovot, Israel, 2013
22- "Bug bytes: bioinformatics for the human microbiome in health and disease," University of Michigan Molecular and Clinical Epidemiology of Infectious Diseases (MAC-EPID) symposium. Ann Arbor, MI, 2013
23- "Computational methods for meta'omic characterization of the human microbiome," Tufts Computer Science Department Colloquium Series. Medford, MA, 2013
24- "Adding depth to human microbiome studies with multi'omic data integration," International Human Microbiome Congress. Hangzhou, China, 2013
25- "Functional analysis of human microbiome metagenomes, metatranscriptomes, and multi'omics," NIH Microbiome Sciences: Vision for the Future workshop. Bethesda, MD, 2013
26- "Bug bytes: bioinformatics for the human microbiome in health and disease," Canadian Student Health Research Forum. Alberta, Canada, 2013
27- "High-precision functional profiling of metagenomes and metatranscriptomes," Enterics Research Investigational Network Cooperative Research Center meeting. Traverse City, MI, 2013
28- "Computational methods for meta'omic characterization of the human microbiome," Los Alamos National Laboratory Center for Nonlinear Studies seminar. Los Alamos, NM, 2013
29- "An introduction to metagenomics," Strategies and Techniques for Analyzing Microbial Population Structure. Woods Hole, MA, 2013
30- "Bug bytes: Computational analysis methods for microbial communities," University of Oregon BioBE center seminar. Eugene, OR, 2013
31- "From microbial surveys to mechanisms of interaction in the human microbiome," University of Colorado at Boulder BioFrontiers Institute seminar. Boulder, CO, 2013
32- "Detailing the human microbiome with meta'omics," New England Primate Research Center. Southboro, MA, 2012

33- "A meta'omic microscope: detailing the human microbiome," Broad Institute annual retreat. Boston, MA, 2012
34- "Computational methods for meta'omic characterization of the human microbiome," Forsyth Institute seminar. Cambridge, MA, 2012
35- "Computational methods for meta'omic characterization of the human microbiome," Procter and Gamble BioFusion Symposium. Cincinnati, OH, 2012

Number of Presentations: 35.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper

TOTAL:

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received		Paper
11/19/2015	8.00	Huang K, Meadow JF, Gevers D, Lemon KP, Bohannan BJ, Huttenhower C, Franzosa EA. Identifying personal microbiomes using metagenomic codes, proceeding of the national academy of sciences (06 2015)

TOTAL: **1**

Number of Manuscripts:

Books

Received Book

TOTAL:

Received Book Chapter

TOTAL:

Patents Submitted

Patents Awarded

Awards

1- ISCB Overton Prize (Harvard School of Public Health, 2015)
2- eLife Sponsored Presentation Series early career award (Harvard School of Public Health, 2014)
3- Presidential Early Career Award for Scientists and Engineers (Harvard School of Public Health, 2012)

Graduate Students

NAME	PERCENT_SUPPORTED	Discipline
Emma Schwager	0.50	
Tiffany Hsu	0.50	
FTE Equivalent:	**1.00**	
Total Number:	**2**	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
Eric Franzosa	0.20
FTE Equivalent:	**0.20**
Total Number:	**1**

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Curtis Huttenhower	0.08	
FTE Equivalent:	**0.08**	
Total Number:	**1**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:..... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):..... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields: 0.00

Names of Personnel receiving masters degrees

NAME	
Oh, Keunyoung (Kevin)	
Total Number:	**1**

Names of personnel receiving PHDs

NAME
Total Number:

Names of other research staff

NAME	PERCENT_SUPPORTED
Afrah Shafquat	0.33
George Weingart	0.31
FTE Equivalent:	**0.64**
Total Number:	**2**

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

See Attachment

Technology Transfer

CCREPE's implementation as an R/Bioconductor package (see http://huttenhower.sph.harvard.edu/ccrepe) has been carried out in collaboration with Weingart Informatics, an independent software development contractor in San Francisco. This has allowed academic development and validation of the algorithm to be carried out efficiently by students and postdoctoral fellows, while Dr. Weingart has provided industry-quality code, unit testing, packaging, documentation, and distribution. He has begun work on a broader software platform for human microbiome analysis, the bioBakery virtual environment, which may be a target for future industry partnership in the lab.

60287-MA: The human microbiome as a multipurpose biomarker

Associate Professor Curtis Huttenhower, Department of Biostatistics, Harvard School of Public Health

Our final technical report for this project includes 1) a method for uniquely identifying human individuals by way of their personal microbial communities using metagenomic codes, 2) a novel Bayesian method (CCREPE) to identify robust correlation and co-exclusion patterns in compositional data, 3) a generative Bayesian model, SparseDOSSA, for generating microbial community data with known covariation and association patterns, and 4) the PICRUSt method and software for metagenome inference from taxonomic profiles using ancestral state reconstruction. Open-source software implementations for all four main results are available at http://huttenhower.sph.harvard.edu/idability, http://huttenhower.sph.harvard.edu/ccrepe, http://huttenhower.sph.harvard.edu/sparsedossa, and http://huttenhower.sph.harvard.edu/picrust, respectively, with additional material at http://huttenhower.sph.harvard.edu/galaxy and http://bitbucket.org/biobakery. Our final publication list includes 14 manuscripts (PMIDs 25964341, 24629344, 23013615, 26157614, 25303518, 22699609, 22699610, 26418763, 25587358, 22904687, 22698087, 24843156, 23975157, and 22807668) and two currently in preparation (CCREPE and SparseDOSSA).

Problem Statement

The human microbiome comprises the communities of microbes carried in and on the body in health and disease, including trillions of bacteria, viruses, archaea, and fungi per individual. These microbiota, which defend us against pathogens and help digest our food, are personalized among individuals. The specific microbes present at any one habitat within an individual become relatively stable during the first several years of life, but change in as-yet-uncharacterized ways as a host is exposed to new environments, diets, locations, and social contacts. The microbial composition of a given individual might thus be linked to his genetic background or early life history, for example, while the metabolism of those microbes would reveal more about his recent medical history or diet. The goals of this project are thus 1) to assess the structure of any microbial habitat and its potential for identifiability (including dietary history, medical history, biometric, demographics or environmental exposures) and 2) to determine the relationships and interactions between the microbial communities within a host.

Results Summary

Identifying personal microbiomes using metagenomic codes

Large-scale investigations of the human microbiome have revealed great variability in the body site-specific taxonomic composition of organisms across healthy individuals. However, it was not previously known whether this variability is sufficiently nonrandom to uniquely identify individuals within a population, nor whether it is also sufficiently stable to continue uniquely identifying individuals over long time periods (weeks, months, or years). We answered these questions by developing a hitting set-based coding algorithm, which defined body site-specific metagenomic codes: sets of microbial taxa or genes prioritized to uniquely and stably identify individuals. Codes capturing strain variation in clade-specific marker genes were able to distinguish among hundreds of individuals at an initial sampling time point. In comparisons with follow-up samples collected 30-300 days later, ~30% of individuals could still be uniquely pinpointed using metagenomic codes from a typical body site.

Codes based on the gut microbiome were exceptionally stable and pinpointed >80% of individuals. The failure of a code to match its owner at a later time point was largely explained

by the loss of specific microbial strains (at current limits of detection) and was only weakly associated with the length of the sampling interval. In addition to highlighting patterns of temporal variation in the ecology of the human microbiome, this work demonstrated the feasibility of microbiome-based identifiability for the first time, a result with important ethical implications for microbiome study design.

In order to construct metagenomic codes that are stable over time, we first identified properties of individual microbial features (OTUs, genes, and genomic regions) that lead to their repeated detection over multiple sampling time points (**Fig. 1**). Features with low prevalence in the population tended to disappear with the passage of time, making them poor choices for maximizing code stability; abundant features were more likely to be stable over time. Trends in prevalence, abundance, and persistence were conserved across different body sites and across features assayed by different technologies, and they were also consistent regardless of whether features were defined as 16S rRNA gene-based OTUs, metagenomic marker gene sequences, or one-kilobase genomic window abundances (genomic regions).

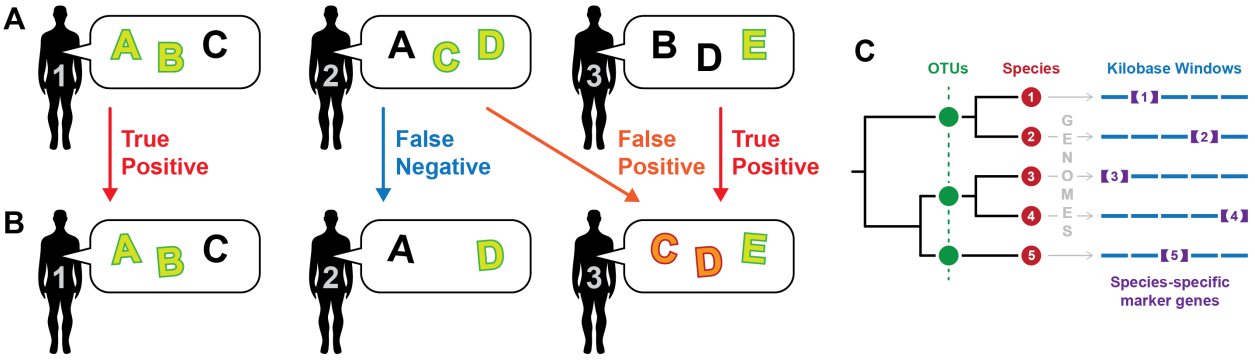


Figure 1: An overview of uniquely identifiable metagenomic code derivation and validation. A) An example of three individuals and their metagenomic features (represented by capital letters) are shown. For each individual, a subset of features is highlighted that is unique among the three individuals. We refer to these sets as metagenomic codes. B) The same three individuals reevaluated after weeks to months. Individual 1's microbiome has remained stable, and his code still uniquely identifies him among the population (a true positive). Individual 2 has lost metagenomic feature C, and his code no longer identifies him (a false negative). Individual 3 has lost feature B and gained feature C. Individual 3 is still a true positive with respect to his own code, but also matches individual 2's code (a false positive). C) Illustration of the four metagenomic feature types considered in our work: OTUs, species, kilobase windows from reference genomes (kbwindows), and species-specific marker genes (markers).

To construct codes for each sample and feature type, we divided all features into three categories within each individual: (i) confidently detected, (ii) confidently absent, and (iii) ambiguous. These categories were defined by a pair of cutoffs that varied by feature type. For example, an OTU was considered confidently detected if its relative abundance met or exceeded 0.001 (0.1%), confidently absent if its relative abundance was below 10-5 (0.001%), and ambiguous otherwise. We defined the features comprising the metagenomic code of an individual to have two critical properties: (i) all code features were confidently detected in that individual at the first sampling time point, and (ii) at least one code feature was confidently absent in each other individual in the population (thus making their ensemble unique to the encoded individual).

To construct a code for an individual X, we first ranked the confidently-detected features of X in order of increasing "abundance gap," which we defined as the abundance of the feature in X minus its next-highest observed abundance in the population. This ranking scheme prioritized the inclusion of abundant features, which were initially determined to be most stable, and penalized the inclusion of features ambiguously detected in other individuals, which would be more likely

to produce false positives at future time points. We then added features from this ranked list to a putative code, at each step "flagging" individuals in the population for whom the next-highest ranked feature was confidently absent (features that would fail to flag new individuals contribute no new information and are skipped). The algorithm terminates when (i) all other individuals have been flagged (i.e. the putative code contains at least one feature that is confidently absent in each other individual, thus making it a unique code) or (ii) we run out of ranked features before flagging all other individuals, in which case we have failed to construct a code for X. Optionally, after assembling a unique code (case i), we can continue adding features to the code until a desired minimum size is reached (7 features in our evaluations). A minimum code size adds robustness to noise and, effectively, error correction to avoid false positives.

Our coding algorithm was able to construct unique codes for almost all individuals by focusing on metagenomically-unique marker genes. These codes were stable in roughly half of the population between the first and second sampling time points (True Positive Rate, TPR = 50%), with relatively low false positive rates. The isolated stool body site was an exception, having a TPR closer to 80%. False negatives were due largely to the disappearance of one or more of the microbial taxa contributing metagenomic features to the code. Notably, the likelihood of a false negative did not appear to depend sensitively on the time sampling interval, as a few weeks to nearly a year were roughly equivalent. Codes based on gene-level features consistently outperformed OTU-based codes. Relative to gene-level codes, OTU-level codes not only depended upon more taxa, but also required the inclusion of less-abundant taxa (which tended to be less stable) to achieve uniqueness. On the other hand, gene-level codes were able to incorporate multiple distinguishing features from the most abundant taxa of an individual, and were therefore more robust to temporal variation. Comparing metagenomic codes to a validation cohort of previously-unseen subjects suggested that codes would tend to remain unique in populations of order-of-magnitude hundreds of individuals.

The results of this study were published in PNAS (PMID 25964341) in collaboration with Brendan Bohannon (University of Oregon) and Katherine Lemon (Forsyth Institute). It was presented at the 2015 Dana-Farber Cancer Institute Biostatistics and Computational Biology seminar series, the 2015 BioC Bioconductor annual meeting, the 2015 Canadian Institute for Health Research International Speaker seminar series, the 2015 Simons Foundation Symposium on Genomics in Single Cells and Microbiomes, the 2015 Harvard CATALYST Understanding Biomarker Science workshop, the 2015 Channing Division of Network Medicine Theodore L. Badger Lecture series, the 2014 University of Oregon Institute for Theoretical Sciences seminar series, the 2014 Statistical and Applied Mathematical Sciences Institute Bioinformatics Opening Workshop, the 2014 META Center for Systems Biology symposium, the 2014 Dalhousie University Centre for Comparative Genomics and Evolutionary Bioinformatics and Microbiome User Group, the 2014 Keystone Symposium on Exploiting and Understanding Chemical Biotransformations in the Human Microbiome, and the 2014 Intelligent Systems for Molecular Biology (ISMB) conference. The project was led by research associate Dr. Eric Franzosa.

Co-occurrence and Co-exclusion Patterns in the Human Microbiome

CCREPE (Compositionality Corrected by REnormalization and PErmutation) has evolved since its inception during this project from an ad hoc approach to a posteriori investigation of associations in microbial community data (initially named ReBoot) to a generalizable Bayesian method for any composition data. Briefly, typical correlation measures such as Pearson or Spearman correlation produce false positives (referred to as spurious correlations) when applied

to compositional data, i.e. measurements in which all values sum to a fixed constant. Proportions, in which fractions sum to 1 or 100%, are a common example; these arise frequently in ecology, since for example microbial taxa are only measurable as counts or as fractions of total community composition. CCREPE corrects the nominal statistical significance of putative correlations in such data in order to report only the significance of association above and beyond that expected due to compositionality alone.

CCREPE's Bayesian model (**Fig. 2**) assumes that a single composition, $\mathbf{C}_i = (C_{i,1}, \ldots, C_{i,p})^T$, is generated by the normalization of a basis, $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,p})^T$. That is,

$$\mathbf{C}_i = \frac{\mathbf{X}_i}{\sum_j X_{i,j}}$$

We also assume that samples are i.i.d., such that $\mathbf{X}_i \overset{iid}{\sim} F_X(\cdot)$ and therefore $\mathbf{C}_i \overset{iid}{\sim} F_C(\cdot)$. Note that $F_C(\cdot)$ is determined from $F_X(\cdot)$ by the transformation from $\mathbf{X}_i$ to $\mathbf{C}_i$ via normalization. The covariance and correlation structures of the basis are denoted by $\mathbf{\Sigma}_X = [\sigma_{X,jj'}]$ and $\mathbf{R}_X = [\rho_{X,jj'}]$, respectively, and the covariance and correlation of the composition by $\mathbf{\Sigma}_C = [\sigma_{C,jj'}]$ and $\mathbf{R}_C = [\rho_{C,jj'}]$. Thus, in order to determine which compositional correlations are significant in the basis, we assess the null hypothesis that $\rho_{X,jj'} = 0$ (or, equivalently, $\sigma_{X,jj'} = 0$) for features j and j', remembering that in real data only the composition and not the basis is observed.

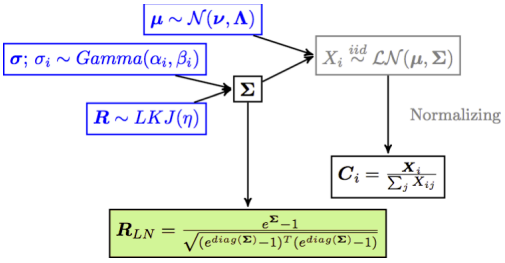


Figure 2: Plate diagram for the CCREPE Bayesian model of significant associations in compositional data. Unobserved correlations Σ in an underlying basis R are generated from mean μ and standard deviation σ, yielding a distribution of true abundances that is assumed to be lognormal (LN). Only the normalized compositions C are observed.

Given that $\mathbf{X} \sim F_X(\cdot)$ with covariance $\mathbf{\Sigma}_X = [\sigma_{X,jj'}]$ and expectation $\boldsymbol{\mu}_X$, a Taylor expansion around $\boldsymbol{\mu}_X$ yields the approximate covariance for $g(\mathbf{X}) = \frac{\mathbf{X}}{\sum_j X_j}$, denoted by $\mathbf{\Sigma}_C = [\sigma_{C,jj'}]$. Defining $\boldsymbol{\omega} = \left(\frac{\mu_{X,1}}{\sum_j \mu_{X,j}}, \ldots, \frac{\mu_{X,p}}{\sum_j \mu_{X,j}}\right)$, this gives:

$$\mathbf{\Sigma}_C = \left(\frac{1}{\sum_j \mu_{X,j}}\right)^2 (\mathbf{I} - \boldsymbol{\omega}\mathbf{1}^T)\mathbf{\Sigma}(\mathbf{I} - \boldsymbol{\omega}\mathbf{1}^T)^T$$

allowing us to predict the behavior of the compositional correlation from the basis parameters that generate it. In particular, for two features in the case where $\sigma_{X,jj'} = 0$ and $j \neq j'$, then:

$$\sigma_{C,jj'} = \left(\frac{1}{\sum_j \mu_{X,j}}\right)^2 \left[\frac{\mu_{X,j}\mu_{X,j'}}{\left(\sum_j \mu_{X,j}\right)^2}\sum_j \sigma_{X,jj} - \frac{\mu_{X,j}}{\sum_j \mu_{X,j}}\sigma_{X,j'j'} - \frac{\mu_{X,j'}}{\sum_j \mu_{X,j}}\sigma_{X,jj}\right]$$

Thus the covariance $\sigma_{C,jj'}$ will be large and positive if both features take up a large proportion of the composition but their variability is small relative to the total variability in the basis. Conversely, covariance will be large and negative if both features take up a large portion of the

composition and a large portion of the total variability, as this arrangement reduces the composition (approximately) to one of two parts. Finally, compositional covariance will also be large and negative if one feature takes up a small portion of the composition with large variability but the other takes up a large portion with small variability, because the feature with large variability forces the other to move in the opposite direction after normalization (**Fig. 3**).

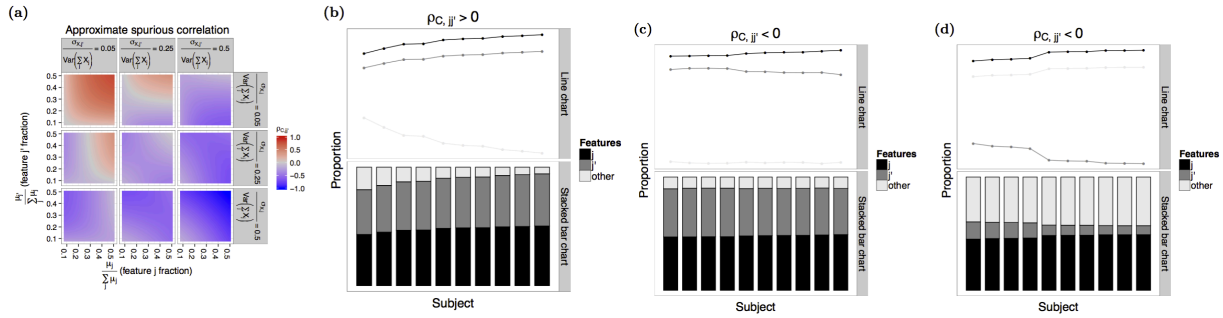


Figure 3: Deriving conditions under which spurious correlations emerge. Three examples that capture the causes of spurious correlation (A) in compositional data between (B) two features of high abundance and high proportional variance, (C) two features of high abundance and low proportional variance, and (D) one feature of high abundance, one low, and high proportional variance. Either high mean μ or standard deviation σ is sufficient, the latter surprisingly so even in cases where feature means are low.

Over the course of developing CCREPE, we also derived a novel ecological similarity measure to use with it, the NC- or N-dimensional Checkerboard score. The NC-score extends the checkerboard score from binary presence-absence variables to ordinal values by re-defining patterns of co-variation and co-exclusion as follows:

1. We define a co-variation pattern as a 2x2 submatrix of the form

$$\{\begin{bmatrix} a & b \\ c & d \end{bmatrix} \mid a < b,\ c < d\ \},$$

or its converse

$$\{\begin{bmatrix} a & b \\ c & d \end{bmatrix} \mid a > b,\ c > d\ \},$$

where a, b, c, $d \in [0, n\text{-}1]$. Biologically, this pattern describes two microbes that co-vary in concert between two samples, i.e. are positively associated.

2. Similarly, we define a co-exclusion pattern as a 2x2 submatrix of the form

$$\{\begin{bmatrix} a & b \\ c & d \end{bmatrix} \mid a > b,\ a > c,\ d > c,\ d > b\ \},$$

or its converse

$$\{\begin{bmatrix} a & b \\ c & d \end{bmatrix} \mid a < b,\ a < c,\ d < c,\ d < b\ \},$$

where a, b, c, $d \in [0, n\text{-}1]$. Biologically, this pattern describes two microbes with different relative abundances between two samples, i.e. negatively associated. In the special binary case of $n = 2$, these two patterns collapse to the standard checkerboard unit. Otherwise, NC-score is equivalent to Kendall's tau calculated on "binned" abundances rather than ranks, which we perform by default using thresholds of "zero" (relative abundance = 0), "very low" (> 0, < 1E-4), "low" (> 1E-4, < .01), "medium" (> .01, < .25), and "high" (> .25, < 1).

Finally, we have applied CCREPE to two relevant microbial datasets, approximately 5,500 16S rRNA gene taxonomic profiles derived from the Human Microbiome Project as published in 2012, and to approximately 2,000 metagenomic species-level taxonomic profiles derived from newly sequenced metagenomes spanning the same subjects analyzed using MetaPhlAn2. The former, published as PMID 22807668, included a global network of 3,005 significant co-occurrence and co-exclusion relationships between 197 clades occurring throughout the human microbiome. This network revealed strong niche specialization, with most microbial associations occurring within body sites and a number of accompanying inter-body site relationships. Microbial communities within the oropharynx grouped into three distinct habitats, which themselves showed no direct influence on the composition of the gut microbiota. Conversely, niches such as the vagina demonstrated little to no decomposition into region-specific interactions. Diverse mechanisms underlay individual interactions, with some such as the co-exclusion of Porphyromonaceae family members and *Streptococcus* in the subgingival plaque supported by known biochemical dependencies. These differences varied among broad phylogenetic groups as well, with the Bacilli and Fusobacteria, for example, both enriched for exclusion of taxa from other clades. Comparing phylogenetic versus functional similarities among bacteria, dominant commensal taxa (such as Prevotellaceae and *Bacteroides* in the gut) often competed, while potential pathogens (e.g. *Treponema* and *Prevotella* in the dental plaque) are more likely to co-occur in complementary niches.

With the latest Bayesian model for CCREPE, we assessed species-level profiles from ~2,000 metagenomes (**Fig. 4**). This yielded a much smaller network of 104 within- and between-site associations, almost all within-site, spanning 115 site-specific clades. Overall, these recapitulated the basic characteristics of earlier 16S-based networks, including little between-site interaction and few "hub" microbes (scale-freeness). We also observed a cluster of co-variation in stool, which contained microbes known to be involved in the health of the gut microbiome, including Escherichia coli and several clades IV and XIVa Clostridia. These new methods will allow the derivation of significant co-variation networks from high-dimensional compositional data, particularly the detection of species and, eventually, sub-species level ecological interactions within the human microbiome and other microbial communities.

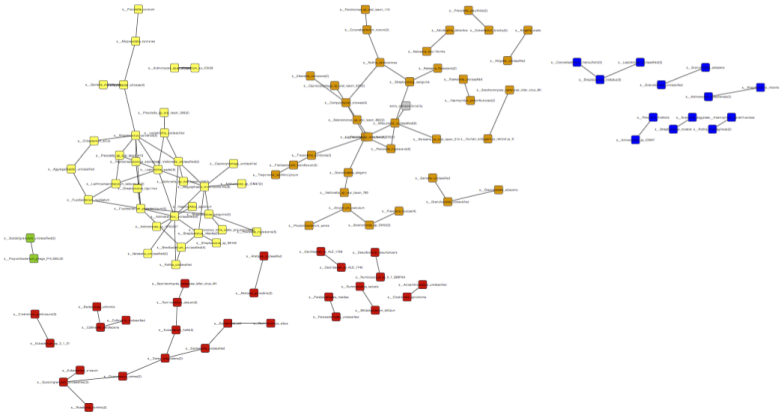


Figure 4: Species-level microbial associations significant by CCREPE in ~2,000 HMP1-II metagenomes. Since publication of approximately 700 body-wide metagenomes during the Human Microbiome Project, an additional ~1,300 metagenomes have been sequenced spanning 100 individuals at three time points each. CCREPE was run on the residuals of a random effects model accounting for repeated measures to identify significant associations between taxa within and across body sites.

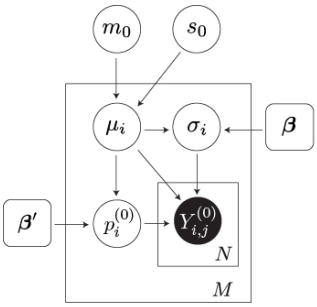
The manuscript describing CCREPE is currently in preparation by lead Ph.D. student Emma Schwager, and during the course of the project period it has included contributions from software developer Dr. George Weingart and M.S. student Craig Bielski. CCREPE has been presented at the 2014 Intelligent Systems for Molecular Biology (ISMB) conference, the 2014 Program in Quantitative Genomics conference, the 2014 Statistical and Applied Mathematical Sciences

Institute Program on Beyond Bioinformatics: Statistical and Mathematical Challenges, the 2014 University of Oregon Institute for Theoretical Sciences seminar series, the 2014 META Center for Systems Biology symposium, the 2014 University of Washington Genome Sciences seminar series, the 2014 Human Microbiome Project consortium meeting, the 2013 Weizmann Institute Systems Biology program seminar, the 2013 Channing Department of Network Medicine seminar series, the 2013 Harvard Medical School Systems Biology program seminar series, the 2013 Tufts University Computer Science department colloquium, the 2013 Human Microbiome Science: Vision for the Future NIH workshop, the 2013 General Meeting of the American Society of Microbiology, the 2013 Harvard School of Public Health Bioinformatics Core forum, the 2013 Los Alamos National Laboratory seminar series, and the 2012 Harvard Medical School Systems Biology program. In addition, CCREPE has been an analysis methodology in four published studies (PMIDs 22807668, 24629344, 24843156, 22699609) and the American Gut study currently in review.

A Hierarchical Probabilistic Model of Microbial Community Structure: Sparse Data Observations for the Simulation of Synthetic Abundances (sparseDOSSA)

While many statistical methods have been developed to facilitate analysis of metagenomic data, to date there have been few efforts to benchmark these methods in an accurate and systematic manner - a critical challenge to the developers of these methods as well as the methods' end-users. To address this and to provide a generalizable model of microbial community profiles, we developed SparseDOSSA (Sparse Data Observations for the Simulation of Synthetic Abundances): a hierarchical model of microbial ecological population structure. SparseDOSSA is capable of simulating realistic metagenomic data with known correlation structures and thus provides a gold standard to enable benchmarking of statistical metagenomics methods.

SparseDOSSA's model captures the marginal distribution of each microbial feature as a truncated, zero-inflated log-normal distribution, with parameters derived in turn from a parent log-normal distribution (**Fig. 5**). The model can be effectively fit to reference microbial datasets in order to parameterize their microbes and communities, or to simulate synthetic datasets of similar population structure (including the optional addition of known correlation structures). We demonstrated SparseDOSSA's utility in three applications: 1) accurately modeling microbial community diversity profiles using a small number of fit parameters trained on baseline (healthy) human microbial communities and communities from inflammatory bowel disease (IBD) patients; 2) generating synthetic communities with simulated environmental and ecological correlation structures; and 3) recapitulating the results of an earlier clustering analysis describing microbial response to diet type in mice. These applications comprise the most common downstream analyses applied in metagenomic sequencing experiments, and thus demonstrate SparseDOSSA's utility as a general-purpose aid for evaluating statistical methods in microbial community analysis.



$$\mu_i \overset{i.i.d}{\sim} LN(m_0, s_0)$$
$$\log(\sigma_i) = \beta_0 + \beta_1 \log(\mu_i)$$
$$\text{logit}(p_i^{(0)}) = \beta_0' + \beta_1' \log(\mu_i)$$
$$Y_{i,j}^{(0)} \overset{i.i.d.}{\sim} p_i^{(0)}\delta_0 + (1 - p_i^{(0)})\lfloor LN(\mu_i, \sigma_i)\rfloor$$

Figure 5: SparseDOSSA provides a generative hierarchical Bayesian model for microbial community taxonomic profiles. This comprises a two-layer structure in which each layer is controlled by a lognormal distribution. Individual microbial features (second layer) are assumed to be drawn from a lognormal distribution with marginal mean μ_i and standard deviation σ_i. In combination with feature-specific sparsity (i.e. expected fraction of zeros) p_i, $Y_{i,j}$ is generated as the number of reads of feature i in sample j. To generate the overall distribution of the population of microbial features, β,β' are parameters connecting the marginal mean parameters to the marginal standard deviation parameters and marginal sparsities, respectively, and μ_0, s_0 are the mean and standard deviation of the overall lognormal distribution (first layer). Rectangles denote replication of the model within the rectangle, i.e. plate structure, with the number of replication is labeled at the bottom-right corner.

We validated SparseDOSSA's ability to accurately fit and simulate microbial communities with ecological patterns recapitulating those in a variety of real community types, by comparing the distribution of microbe-specific means and beta-diversity profile in simulated data to the real data that are used to train the model. This ultimately allows us and others to quantitatively benchmark analysis methods for microbiome features, their ecological properties, and their associations with environment and host phenotypes. We first assessed the degree to which SparseDOSSA's fitted two-layer model captured the marginal variation of microbial community taxonomic profiling data across all microbes. To that end, we first compared the trend of rank average abundance of all microbes in two real datasets (Morgan Genome Bio 2012, 250 samples and 158 genera, and HMP Nature 2012, 16S posterior fornix subset) to those of simulated datasets generated by the model fitted on these real datasets. The fitting was repeated using both the naïve and fully Bayesian fitting procedures, with simulated datasets generated based on SparseDOSSA's two-layer model setting parameters to be the estimates from the fitting step and with the same dimension as the real dataset. This verified the model's ability to capture the variation pattern of marginal microbial community profiles and the validity of our naive fitting method as a fast approximation of the full Bayesian fitting method (**Fig. 6**).

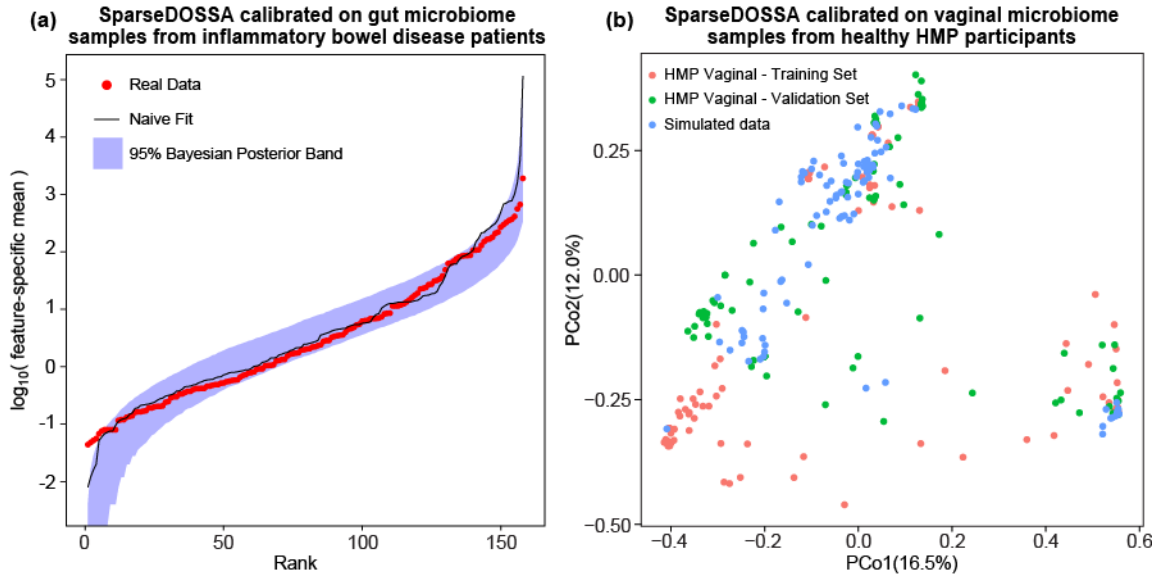


Figure 6: The SparseDOSSA model accurately captures feature mean distributions and beta diversities of microbial communities. A) Rank abundances of log-transformed feature specific means for data from Morgan et al

and from SparseDOSSA simulated data based on this, using naive estimates and the pointwise 95% posterior interval derived from fully Bayesian fitting of the same model. B) MDS analysis on Bray-Curtis distances of taxonomic profiles from 190 vaginal microbial communities in the Human Microbiome Project split into a training subset (95 samples) for model fitting, a validation set (95 samples), and 95 simulated samples from the resulting SparseDOSSA model (all containing 172 microbial features).

When used as a model for simulating realistic microbial community data, one major function of SparseDOSSA is to impose a known "phenotypic response" in synthetic microbial features. This is done by capturing the overall diversity pattern of a community and subsequently also inducing artificial correlations between features and sample properties such as environment or phenotype. We verified the model's ability to include detectable categorical feature-metadata associations in its simulated output by artificially associating nine randomly chosen features and a binary environmental metadatum with 250 samples and the Morgan et al dataset as template. All synthetic associations were detected among the 17 features of greatest effect size, with no false negatives. Interestingly, several apparent false positives (i.e. features that are spuriously differentially abundant) occur at this level of significance, caused not by the SparseDOSSA model but by the known effects of compositionality (i.e. relative abundance normalization) in ecological data. Simulated associations with a synthetic continuous metadatum were similarly successful, including the model's ability to correctly capture sparsity patterns in microbial features (**Fig. 7**).

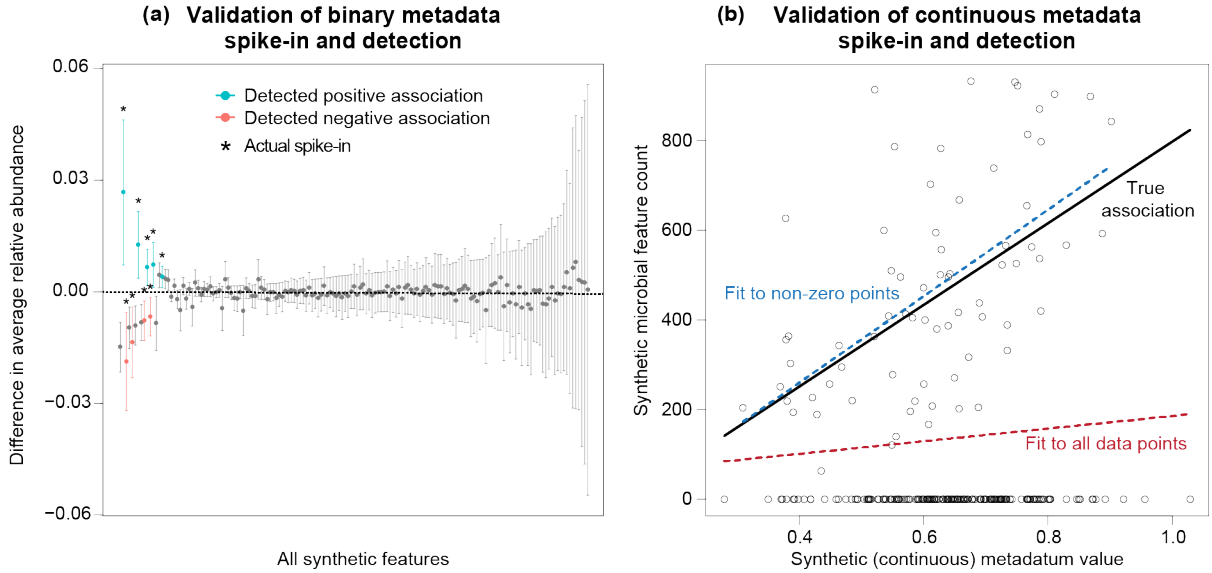


Figure 7: Simulating categorical or continuously valued population variability among microbial community samples. A) Differences of mean relative abundances between two classes of a simulated binary metadatum based on Morgan et al, along with the empirical inter-quatile range of all features as contrasted between metadatum levels. B) Correlation of one feature into which an association to a continuously variable metadatum has been spiked (Y axis) with that metadatum's value (X axis).

The SparseDOSSA manuscript is currently under review as written by project lead Ph.D. student Boyu Ren and research scientist Dr. Eric Franzosa, with contributions over the course of the project from undergraduate research assistants Joseph Moon and Yiren Lu. SparseDOSSA has been presented at the 2015 Dana-Farber Cancer Institute Biostatistics and Computational Biology seminar series and at the 2014 Statistical and Applied Mathematical Sciences Institute Program on Beyond Bioinformatics.

Inferring microbial community metagenomes from marker gene sequences

Profiling phylogenetic marker genes, such as the 16S rRNA gene, is a key tool for studies of microbial communities but does not provide direct evidence of a community's functional capabilities. We developed PICRUSt (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States), a computational approach to predict the functional composition of a metagenome using marker gene data and a database of reference genomes. PICRUSt uses an extended ancestral-state reconstruction algorithm to predict which gene families are present and then combines gene families to estimate the composite metagenome. Using 16S information alone, PICRUSt recaptures key findings from the Human Microbiome Project and accurately predicts the abundance of gene families in host-associated and environmental communities, with quantifiable uncertainty. Phylogeny and function are thus sufficiently linked that this 'predictive metagenomic' approach has begun to provide useful insights into the thousands of uncultivated microbial communities for which only marker gene surveys are currently available (**Fig. 8**).

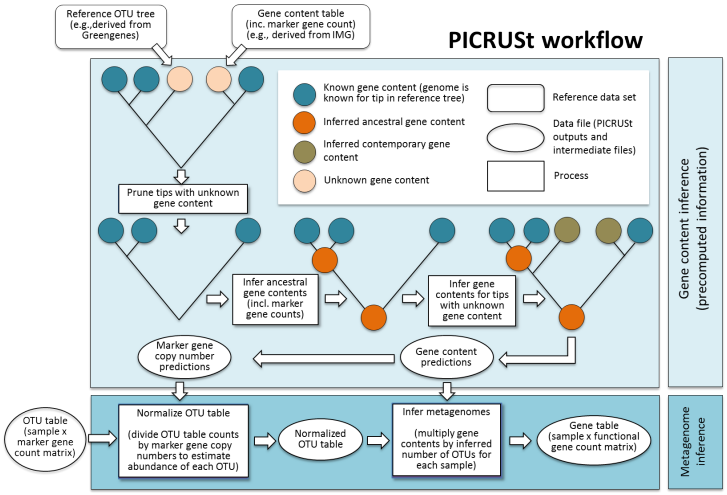


Figure 8: The PICRUSt workflow. PICRUSt is composed of two high-level workflows: gene content inference (top) and metagenome inference (bottom). Beginning with a reference phylogeny and a gene content table, gene content inference predicts genes for each taxon with unknown content, including marker gene copy. Metagenome inference takes a taxonomic profile, where taxa correspond to tips in the reference tree, as well as the copy number of the marker gene in each taxon and its gene content and outputs a metagenome table on a per-sample basis.

We developed PICRUSt to predict the functional composition of a microbial community's metagenome from its 16S profile. This is a two-step process. In the initial 'gene content inference' step, gene content is precomputed for each organism in a reference phylogenetic tree. This reconstructs a table of predicted gene family abundances for each organism (tip) in the 16S-based phylogeny. Because this step is independent of any particular microbial community sample, it is pre-calculated only once. The subsequent 'metagenome inference' step combines the resulting gene content predictions for all microbial taxa with the relative abundance of 16S rRNA genes in one or more microbial community samples, corrected for expected 16S rRNA gene copy number, to generate the expected abundances of gene families in the entire community.

The value of PICRUSt depends on the accuracy of its predicted metagenomes from marker gene samples and the corresponding ability to recapitulate findings from metagenomic studies. The performance of PICRUSt was first evaluated using the set of 530 HMP samples that were analyzed using both 16S rRNA gene and shotgun metagenome sequencing. We treated HMP metagenomic samples as a reference and calculating the correlation of PICRUSt predictions from paired 16S samples across 6,885 resulting orthologous groups. Predictions had high agreement with metagenome sample abundances across all body sites (Spearman $r=0.82$, $p<0.001$). As a targeted example, we also tested PICRUSt's accuracy in specifically predicting the abundance of

glycosaminoglycan (GAG) degradation functions, which are more abundant in the gut than elsewhere in the body. Using the same differential enrichment analysis on both PICRUSt and metagenomic data yielded identical rankings across body sites and very similar quantitative results, suggesting that PICRUSt predictions can be used to infer biologically meaningful differences in functional abundance from 16S surveys even in the absence of comprehensive metagenomic sequencing.

Next, we then evaluated the prediction accuracy of PICRUSt in metagenomic samples from a broader range of habitats including mammalian guts, soils from diverse geographic locations, and a phylogenetically complex hypersaline mat community (**Fig. 9**). These habitats represent more challenging validations than the human microbiome, as they have not generally been targeted for intensive reference genome sequencing. Because PICRUSt benefits from reference genomes that are phylogenetically similar to those represented in a community, this evaluation allowed us to quantify the impact of increasing dissimilarity between reference genomes and the metagenome.

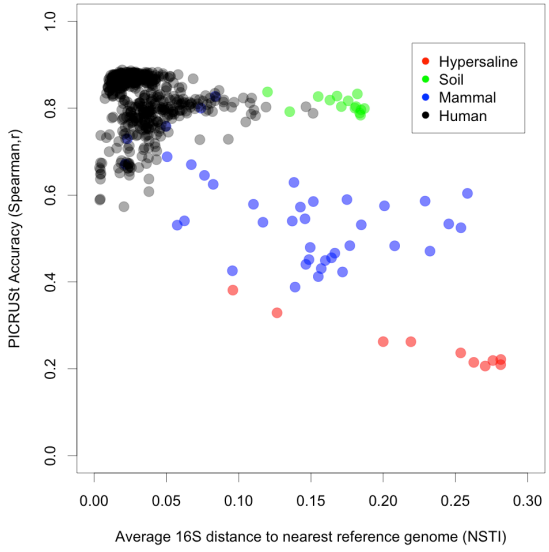


Figure 9: PICRUSt accuracy across various environmental microbiomes. Prediction accuracy for paired 16S rRNA marker gene surveys and shotgun metagenomes (y-axis) are plotted against the availability of reference genomes as summarized by the Nearest Sequenced Taxon Index (NSTI; x-axis). Accuracy is summarized using the Spearman correlation between the relative abundance of gene copy number predicted from 16S data using PICRUSt versus the relative abundance observed in the sequenced shotgun metagenome. In the absence of large differences in metagenomic sequencing depth (see text), relatively well-characterized environments, such as the human gut, have low NSTI values and can be predicted accurately from 16S surveys. Conversely, environments containing much unexplored diversity (e.g. phyla with few or no sequenced genomes), such as the Guerrero Negro hypersaline microbial mats, tended to have high NSTI values.

To characterize this effect, we developed the Nearest Sequenced Taxon Index (NSTI) to quantify the availability of nearby genome representatives for each microbiome sample. NSTI is the sum of phylogenetic distances for each organism in the taxonomic profile to its nearest sequenced reference genome, measured in terms of substitutions per site in the marker (i.e. 16S rRNA) gene and weighted by the frequency of that organism in the profile. As expected, NSTI values were greatest for the phylogenetically diverse hypersaline mat microbiome (mean NSTI=0.23 +/- 0.07 s.d.), least for the well-covered HMP samples (mean NSTI=0.03 +/- 0.02 s.d.), mid-range for the soils (mean NSTI=0.17 +/- 0.02 s.d.) and varied for the mammals (mean NSTI=0.14 +/- 0.06 s.d.). Also as expected, the accuracy of PICRUSt in general decreased with increasing NSTI across all samples (Spearman $r=-0.4$, $p< 0.001$) and within each microbiome type (Spearman $r=-0.25$ to -0.82, $p<0.05$). For a subset of mammal gut samples (NSTI<0.05) and all of the soil samples that we tested, PICRUSt produced accurate metagenome predictions (Spearman $r=0.72$ and 0.81, respectively, both $p<0.001$). Both the mammal and hypersaline metagenomes were shallowly sequenced at a depth expected to be insufficient to fully sample the underlying community's genomic composition, thus likely causing the accuracy of PICRUSt to appear artificially lower for these communities. Although the lower accuracy on the hypersaline

microbial mats community (Spearman r =0.25, p<0.001) confirms that PICRUSt must be applied with caution to the most novel and diverse communities, the ability to calculate NSTI values within PICRUSt from 16S data allows users to determine whether their samples are tractable for PICRUSt prediction prior to running an analysis. Moreover, the evaluation results verify that PICRUSt provides useful functional predictions for a broad range of environments beyond the well-studied human microbiome.

PICRUSt was published in Nature Biotechnology (PMID 23975157) in 2013 and has been cited approximately 150 times since then, thanks to lead authors postdoctoral fellows Drs. Morgan Langille and Jesse Zaneveld, with contributions over the course of the project from Ph.D. student Joshua Reyes and collaborators Drs. Rob Knight (UCSD) and Rob Beiko (Dalhousie). It has been presented in a wide variety of venues including:

- "From microbes to molecules: detailing function in integrated multi'omics." New York Academy of Sciences Advances in Human Microbiome Science. New York, NY, 2015
- "High-precision functional profiling of microbial communities and the human microbiome." Wellcome Trust Workshop on Applied Bioinformatics and Public Health Microbiology. Hinxton, UK, 2015
- "High-precision functional profiling of microbial communities and the human microbiome." Canadian Institute for Health Research STAGE seminar. Toronto, Canada, 2015
- "Towards systems-level functional profiling of microbial communities and the human microbiome." University of Pennsylvania Microbiology seminar. Philadelphia, PA, 2015
- "High-precision Functional Profiling of Microbial Communities and the Human Microbiome." 41st Annual Northeast Bioengineering Conference. Albany, NY, 2015
- "High-precision functional profiling and integration of metagenomes and metatranscriptomes." International Human Microbiome Congress workshop on Integrated 'Omics for Microbiome Analyses. Luxembourg, Luxembourg, 2015
- "High-precision functional profiling of microbial communities and the human microbiome." Simons Foundation Symposium on Genomics in Single Cells and Microbiomes. New York, NY, 2015
- "A Tour of the bioBakery: Computational Tools for Microbial Community Analysis." Broad Institute Medical and Population Genetics seminar. Cambridge, MA, 2015 (presented by Eric Franzosa)
- "Towards systems-level functional profiling of microbial communities and the human microbiome." Channing Division of Network Medicine Theodore L. Badger Lecture. Boston, MA, 2015
- "The microbiome in IBD and analysis methods for microbial communities." International Inflammatory Bowel Disease Genetics Consortium meeting. Barcelona, Spain, 2015
- "An Introduction to Microbial Community Analyses." Evomics and Genomics workshop. Cesky Krumlov, Czech Republic, 2015
- "High-specificity methods for profiling microbial communities and the human microbiome." University of Oregon Computer Science colloquium. Eugene, OR, 2014
- "Metagenomics, metatranscriptomics, and multi'omic integration." Massachusetts General Hospital Center for the Study of Inflammatory Bowel Disease research symposium. Boston, MA, 2014
- "High-precision methods for metagenomic and metatranscriptomic profiling." New York University Medical School seminar. New York, NY, 2014

- "High-precision profiling of microbial communities and the human microbiome." University of Oregon Institute for Theoretical Sciences seminar. Eugene, OR, 2014
- "Computational Approaches for the Human Microbiome in Health and Disease," 12th Biennial Congress of the Anaerobe Society of the Americas. Chicago, IL, 2014
- "An introduction to the microbiome and quantitative methods for microbial community analysis," HSPH Biostatistics Summer Program in Quantitative Sciences. Boston, MA, 2014
- "An introduction to the microbiome and methods for microbial community analysis," Harvard/MIT Minority Introduction to Engineering, and Science. Boston, MA, 2014
- "An introduction to metagenomics," Strategies and Techniques for Analyzing Microbial Population Structure. Woods Hole, MA, 2014
- "Identifiability of the Human Microbiome," Dalhousie University Centre for Comparative Genomics and Evolutionary Bioinformatics and Microbiome User Group. Halifax, Canada, 2014
- "A Tour of the BioBakery: Computational Tools for Microbial Community Analysis," Harvard School of Public Health Program in Quantitative Genomics Short Course series. Boston, MA, 2014 (presented by Eric Franzosa)
- "High-precision functional profiling and integration of metagenomes and metatranscriptomes," Weizmann Institute Systems Biology Seminar Series. Rehovot, Israel, 2013
- "Bug bytes: bioinformatics for the human microbiome in health and disease," University of Michigan Molecular and Clinical Epidemiology of Infectious Diseases (MAC-EPID) symposium. Ann Arbor, MI, 2013
- "Computational methods for meta'omic characterization of the human microbiome," Tufts Computer Science Department Colloquium Series. Medford, MA, 2013
- "Adding depth to human microbiome studies with multi'omic data integration," International Human Microbiome Congress. Hangzhou, China, 2013
- "Functional analysis of human microbiome metagenomes, metatranscriptomes, and multi'omics," NIH Microbiome Sciences: Vision for the Future workshop. Bethesda, MD, 2013
- "Bug bytes: bioinformatics for the human microbiome in health and disease," Canadian Student Health Research Forum. Alberta, Canada, 2013
- "High-precision functional profiling of metagenomes and metatranscriptomes," Enterics Research Investigational Network Cooperative Research Center meeting. Traverse City, MI, 2013
- "Computational methods for meta'omic characterization of the human microbiome," Los Alamos National Laboratory Center for Nonlinear Studies seminar. Los Alamos, NM, 2013
- "An introduction to metagenomics," Strategies and Techniques for Analyzing Microbial Population Structure. Woods Hole, MA, 2013
- "Bug bytes: Computational analysis methods for microbial communities," University of Oregon BioBE center seminar. Eugene, OR, 2013
- "From microbial surveys to mechanisms of interaction in the human microbiome," University of Colorado at Boulder BioFrontiers Institute seminar. Boulder, CO, 2013
- "Detailing the human microbiome with meta'omics," New England Primate Research Center. Southboro, MA, 2012

- "A meta'omic microscope: detailing the human microbiome," Broad Institute annual retreat. Boston, MA, 2012
- "Computational methods for meta'omic characterization of the human microbiome," Forsyth Institute seminar. Cambridge, MA, 2012
- "Computational methods for meta'omic characterization of the human microbiome," Procter and Gamble BioFusion Symposium. Cincinnati, OH, 2012